Artificial Intelligence in Next Generation Sequencing for Oncology: A National Framework for Clinical Adoption

ISBN: 978-1-943295-26-5

Uma Nambiar
Sriram Menon Koottala
Maria Martin
Gopika K
TATA IISc Medical School Foundation
(umanambiar@iiscmedicalschoolfoundation.org)
(junior.executive@iiscmedicalschoolfoundation.org)
(intern7@iiscmedicalschoolfoundation.org)
(Intern6@iiscmedicalschoolfoundation.org)

Artificial Intelligence improves Next Generation Sequencing by automating the detection and interpretation of genetic variants, helping oncologists identify specific mutations that guide cancer treatment. However, reliance on algorithms without proper validation can lead to misclassification and ethical concerns related to data use. This paper promotes a dual validation process, in which AI performs the first level analysis and trained molecular pathologists confirm the results according to recognised laboratory standards such as ISO 15189 2012 (NABL). Managed responsibly, this collaboration between AI and human expertise strengthens accuracy, patient safety, and trust in genomic medicine.

Keywords: Artificial Intelligence, Next Generation Sequencing, Oncology, Precision Medicine

1. Introduction

Next Generation Sequencing has fundamentally changed the way oncologists diagnose and manage cancer. Instead of analysing a few genes, NGS enables the simultaneous sequencing of hundreds or thousands of genes, providing a panoramic view of a tumor's genetic architecture. This information supports decisions about targeted therapies, immunotherapy, and prognostic classification.

The challenge lies in interpreting this vast quantity of raw sequencing data. A single tumor sample may generate more than one hundred gigabytes of information, containing millions of short DNA fragments that must be aligned to a reference genome and evaluated for authenticity. Manual review of this data is slow and prone to inconsistency. Artificial Intelligence provides the computational capacity and pattern recognition capability necessary to process this information reliably.

Machine learning algorithms trained on validated genomic datasets can identify variants that might otherwise be overlooked, detect sequencing artefacts, and classify mutations according to their likely clinical impact. By linking genomic information with clinical outcomes, Artificial Intelligence has begun to close the gap between data acquisition and clinical insight. For India, with its large population and diverse genetic backgrounds, these technologies represent an opportunity to democratize advanced molecular diagnostics provided they are implemented responsibly.

2. The Evolution of Artificial Intelligence in Variant Calling

Traditional variant callers used deterministic rules and Bayesian statistics to evaluate whether a nucleotide change represented a real mutation or a sequencing error. This approach was effective for high quality data but struggled with low coverage regions or complex variants. Deep learning models changed this paradigm by learning directly from the data rather than relying on predefined mathematical assumptions.

Deep Variant, developed by researchers at Google, translates sequencing reads into image-like matrices and uses convolutional neural networks originally designed for computer vision to classify each potential variant. The system can detect subtle signal differences between true variants and background noise, improving accuracy for both short read and long read platforms.

Clair and **Medaka**, created by Oxford Nanopore Technologies, address the higher error rates characteristic of nanopore sequencing. They employ recurrent neural networks capable of processing long sequential data to identify base modifications and small insertions or deletions. These tools make real time analysis feasible, enabling on site sequencing in oncology wards or field hospitals.

Deep Somatic, published in Nature Biotechnology, extends these concepts to the somatic mutation landscape of tumors. Its architecture uses residual networks that integrate read depth, strand bias, and contextual sequence features, producing accurate calls for single nucleotide variants and small indels across multiple cancer types.

These systems consistently outperform traditional pipelines such as GATK in sensitivity and precision. However, every algorithm must undergo extensive validation against reference genomes and clinical controls before use in patient reporting, consistent with ISO 15189 requirements for analytical verification.

3. The DRAGEN Pipeline and Pangenome Mapping

The DRAGEN platform, described in Nature Biotechnology (2024), represents one of the most mature Artificial Intelligence

enabled sequencing frameworks. It integrates base calling, alignment, and variant detection into a single hardware accelerated system using Field Programmable Gate Arrays. This approach provides both speed and reproducibility.

A notable innovation of DRAGEN is its support for pangenome references. Instead of aligning reads to a single linear genome, it aligns them to a graph based structure that contains genetic diversity from multiple populations. This eliminates the reference bias that can cause systematic under detection of population specific variants.

For India, where population heterogeneity is among the highest in the world, pangenome mapping is vital. Integration of Indian genome data into these algorithms can significantly enhance variant discovery in under represented ethnic groups. Shared national bioinformatics infrastructure could host DRAGEN or comparable pipelines centrally, allowing smaller hospitals to access high accuracy variant calling without building their own data centers.

4. Integrating Multi Omics and Clinical Data

Modern oncology demands more than genomic data alone. Transcriptomics, proteomics, metabolomics, and imaging information provide complementary insights into tumor biology. Artificial Intelligence can combine these data layers into a unified predictive framework.

A 2025 NPJ Digital Medicine review demonstrated how neural networks trained on multi omics datasets can correlate specific genomic mutations with downstream molecular pathways and clinical outcomes. Such systems can predict drug response, identify potential resistance mechanisms, and stratify patients for clinical trials.

For example, integrating gene expression signatures with NGS derived mutation profiles can predict immunotherapy response in lung and melanoma patients. In breast cancer, combining mutation data with proteomic profiles improves subtype classification beyond conventional hormone receptor testing.

Realizing this potential in India requires interoperable health information systems that can link laboratory, imaging, and clinical data under secure governance. The National Digital Health Mission's architecture provides a foundation for this integration, provided genomic privacy and consent frameworks are fully developed.

5. Clinical and Regulatory Frameworks

International regulatory bodies emphasize that Artificial Intelligence should augment, not replace, human expertise. The 2024 European Society for Medical Oncology recommendations outlines clear guardrails:

- Each laboratory must validate its bioinformatics workflow
- Document algorithm performance
- Retain clinician oversight for all interpretive decisions.

Similarly, the United States Food and Drug Administration have provided detailed documentation for cleared Artificial Intelligence assisted in vitro diagnostic sequencing systems. These reports specify analytical performance, reference dataset composition, and post market monitoring obligations.

For Indian laboratories, the logical path is alignment with ISO 15189 and NABL accreditation while incorporating ESMO and FDA best practices. This includes,

- Establishing standard operating procedures for algorithm updates
- Maintaining version control
- Auditing Artificial Intelligence outputs through human verification.

Such governance will ensure that Indian laboratories can participate confidently in international clinical research and therapeutic trials.

6. Operational Considerations for Indian Laboratories

The successful implementation of Artificial Intelligence in sequencing depends as much on operations as on algorithms.

- **Infrastructure:** Deep learning pipelines require robust computing resources, stable power, and high speed data storage. Laboratories must invest in redundant systems and secure data networks that comply with health information privacy standards.
- **Workforce:** Clinicians, pathologists, and technologists need structured training in data interpretation and Artificial Intelligence literacy. Understanding confidence scores, variant classification standards, and limitations of automation is essential to avoid over reliance on algorithms.
- **Patient Consent:** Sequencing data can reveal hereditary risk and familial relationships. Consent procedures must explicitly inform patients about data storage, sharing, and possible secondary use.
- **Collaborative Roles**: New positions such as Genomic Data Scientist, Bioinformatics Coordinator, and Artificial Intelligence Validation Officer can enhance accountability. Assigning these roles elevates technical staff morale and transforms the laboratory into a multidisciplinary innovation hub.
- **Sustainability:** Smaller centers can access Artificial Intelligence analysis through shared cloud nodes managed by accredited hubs. This model reduces cost while maintaining standardization and traceability.

7. Ethical and Social Challenges

Artificial Intelligence is only as unbiased as its training data. Most publicly available genomic datasets represent European

and North American populations. Applying such models in India without local retraining risks systematic inaccuracies. Developing a national genomic database representing India's genetic diversity is therefore a strategic priority.

Algorithmic transparency is equally important. Laboratories must be able to trace each decision made by Artificial Intelligence systems and verify the evidence behind every variant call. Black box systems that cannot provide explainable reasoning should not be used for clinical reporting.

Patient trust depends on open communication about how Artificial Intelligence contributes to diagnosis. Clinicians must remain the ultimate decision makers, interpreting results within the full clinical context.

8. Education, Learning Curve, and Capacity Building

Introducing Artificial Intelligence into clinical sequencing requires a cultural shift within the medical community. Training programs should combine molecular pathology, bioinformatics, and data ethics. Universities and professional societies can collaborate to develop certification courses for Artificial Intelligence proficiency in laboratory medicine.

For clinicians, understanding Artificial Intelligence generated confidence intervals, probability distributions, and variant classifications are critical. Without such literacy, there is a risk of misinterpretation or misplaced confidence in automated reports. Hospitals can introduce regular Artificial Intelligence case discussions similar to tumor boards, allowing doctors and scientists to review model outputs collectively.

Continuous education will reduce apprehension, improve collaboration, and ensure that Artificial Intelligence adoption enhances rather than disrupts clinical workflow.

9. Future Outlook

Artificial Intelligence in sequencing is evolving toward explainable and adaptive systems. Future models will not only detect mutations but also generate visual explanations linking each decision to supporting evidence such as read depth and alignment quality. Integrating radiology, pathology, and genomic data within a single analytical environment will enable real time clinical decision support.

Regulatory agencies are now defining standards for Artificial Intelligence as a medical device. Aligning Indian practice with these international frameworks will prepare national laboratories for cross border collaborations.

Projects such as Genome India and the Indian Council of Medical Research's molecular oncology initiatives can supply population specific data for model training. This will ensure that Artificial Intelligence tools reflect the genetic diversity and clinical realities of the Indian population, making precision oncology both inclusive and sustainable.

10. Conclusion

Artificial Intelligence has become an essential companion to Next Generation Sequencing in oncology. Deep learning models such as Deep Variant, Clair, Medaka, Deep Somatic, and DRAGEN enhance detection accuracy and reduce reporting time. Multi omics integration transforms isolated genetic data into actionable knowledge that guides treatment decisions.

However, technology must operate within a strong framework of validation, ethics, and human oversight. Laboratories must adhere to ISO 15189 and NABL standards, invest in staff training, and establish transparent governance for data and algorithms.

By combining machine precision with human judgment, India can create a genomic ecosystem that is innovative, ethical, and globally competitive. Artificial Intelligence, when used responsibly, can accelerate the transition from reactive cancer care to truly personalized and preventive oncology.

11. Acknowledgment

The authors acknowledge the Genomics Core Laboratory and Bioinformatics Unit of the TATA IISc Medical School Foundation for their guidance and infrastructure support.

Author Contributions

- Dr. Uma Nambiar conceptualized and supervised the study.
- Dr. Sriram Menon Koottala developed the operational and accreditation framework.
- Ms. Gopika K and Ms. Maria Martin conducted literature synthesis and ensured alignment with ISO 15189 and NABL quality systems.
- All authors reviewed and approved the final manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

12. References

1. NPJ Digital Medicine, Nature (2025). Artificial Intelligence and Machine Learning in Precision Oncology Integrating Multi Omics with Clinical Data.
2. PubMed Central (2025). Review of Artificial Intelligence in Variant Calling using Deep Variant, Clair, and Medaka.

3. Nature Biotechnology (2024). DRAGEN Pipeline using Machine Learning for Variant Detection and Pangenome Mapping.
4. Nature Biotechnology (2025). Deep Somatic Deep Learning for Somatic Single Nucleotide Variants and Indels.
5. Annals of Oncology, ESMO (2024). Clinical Recommendations for Next Generation Sequencing in Advanced Cancer.
6. FDA Access Data. Documentation of Cleared Next Generation Sequencing Cancer Diagnostic Devices and Algorithm Validation Examples.